UCRL-JC-123930 Abs

Building the LLNL/UCSF Digital Mammogram Library with Image Groundtruth

Laura N. Mascio*, Steven D. Frankel, M.D.**, Jose M. Hernandez*, Clint M. Logan*

*Lawrence Livermore National Laboratory, Livermore, CA, USA;
**Department of Radiology, University of California SF, San Francisco, CA, USA
e-mail: lmascio@llnl.gov

Automated computer inspection of mammograms can provide assistance in the early detection of breast cancer. Many researchers are developing algorithms to detect one or more of the discrete lesions: microcalcifications (MCs), spiculated lesions and circumscribed masses. The evaluation and comparison of these algorithms has been difficult, not least because most groups report results based on their own data set. Attempts to compare the quality, difficulty and/or subtlety of data sets are necessarily subjective. Performance metrics for evaluating the algorithms vary and are often subjective due to a lack of objective truth data.

In order to address these issues for mammograms featuring MCs, we developed a 12-volume CD library containing digitized mammogram films, associated groundtruth images at full resolution, clinical history information, and radiologists' comments. In practice, the following issues had to be addressed and resolved: providing adequate mammogram display and drawing capability on a computer monitor; characterizing the films; selecting text information to include with each image or case; choosing the image data file format; and establishing filename conventions.

Methods:

A radiology team selected 198 films from 50 patients (4 films per case, minus 2 films for one case with unilateral mastectomy) so as to span a range of cases of interest: 5 normal, average, healthy cases (previous normal mammograms and no history of ultrasound, magnification views, biopsy, etc.), 5 normal but difficult cases (with either dense or fibrous breasts, implants, or asymetric tissue), 20 cases of obviously benign microcalcifications (with at least 3 years of follow-up without change or developing cancer), 12 cases of suspicious, benign microcalcifications, (note: all these benign cases had either a biopsy or a diagnostic mammogram plus at least 3 years of subsequent follow-up without change or developing cancer), and 8 malignant clustered MC cases, biopsy proven.

The selected films were digitized using the Dupont NDT 35 micron film digitizer (designed for precision industrial radiography) using the Fairchild anti-blooming CCD chip. The pixels are square, with a range of 4096 grayscale levels. The field of view covers an area with maximum dimensions up to 7x17 inches, or 18x43 centimeters. The resulting digital mammograms are on the order of 40 megabytes each.

To display and annotate these images, a custom drawing package was required. It handles digital image presentation, and allows annotation at full resolution for maximum drawing accuracy.

Annotation included marking the extent of some exemplary individual calcifications (from isolated, benign MCs to individual members of a cluster), as well as outlining all of the calcification cluster regions. The markings were made while referring to all available clinical information, including the original x-rays and years of follow-up; where applicable, results from ultrasound, spot compression magnification projections, and biopsy were used.

Simultaneously, information for each film was entered into a database, (using the BI-RADS terminology):
a) case history, and clinical diagnosis

b) subjective estimation of gladularity ("density") of the breast (on a scale of 0 to 3) as viewed on the film,
c) subjective estimation of lesion subtlety (on a scale of 0 to 5 as defined by the University of Chicago),
d) estimated 3D quadrant location of indicated lesion and
e) radiologists' comments regarding the case or the annotation.

The digital mammograms and binary mask truth images were stored in the Image Cytometry Standard (ICS) image data format. This format was selected because its size is small and manageable for research purposes, while it has the following desirable properties: the ascii header file can be viewed with any text viewing tool, or editor; the header is extensible (key word accessible) and can be expanded to include patient information, subsequent processing and analysis results, or other data. The data file is a row-order binary stream file that can be easily imported into many viewing and processing packages. The next generation of this format will have inherent lossless compression and will allow related images of different sizes (e.g. four views from one exam) to be stored in one data file.

Conclusion:

A CD Library is now available, which contains per-pixel groundtruth for mammograms featuring calcifications. Also included in the library is patient information, radiologists' comments and radiologists' estimation of the image characteristics.

The CDs were pressed in a standard (High Sierra) format for maximum accessibility, and a data file format (ICS) was chosen for its flexible features and ease of use.

The per-pixel groundtruth images allow the results of computer detection algorithms to be compared directly to the truth image. This provides a platform for establishing standard metrics to evaluate algorithm performance. With standard data and performance metrics, algorithm development and evaluation can be gauged objectively, and the relative strengths of various algorithms can be identified.

*This work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under contract No. W-7405-Eng-48